


NTK

50°6'14.083"N, 14°23'26.365"E

Národní technická knihovna
National Technical Library

Jak dokumentům automaticky přiřadit hesla PSH

Searching Session  4.10. 2011

Jindřich Mynarz

@jindrichmynarz

Ctibor Škuta

@skutac

Tomáš Müller

@maddog_cz

<http://bit.ly/psh-indexace>

věcné pořádání řízený slovník
předmětová hesla

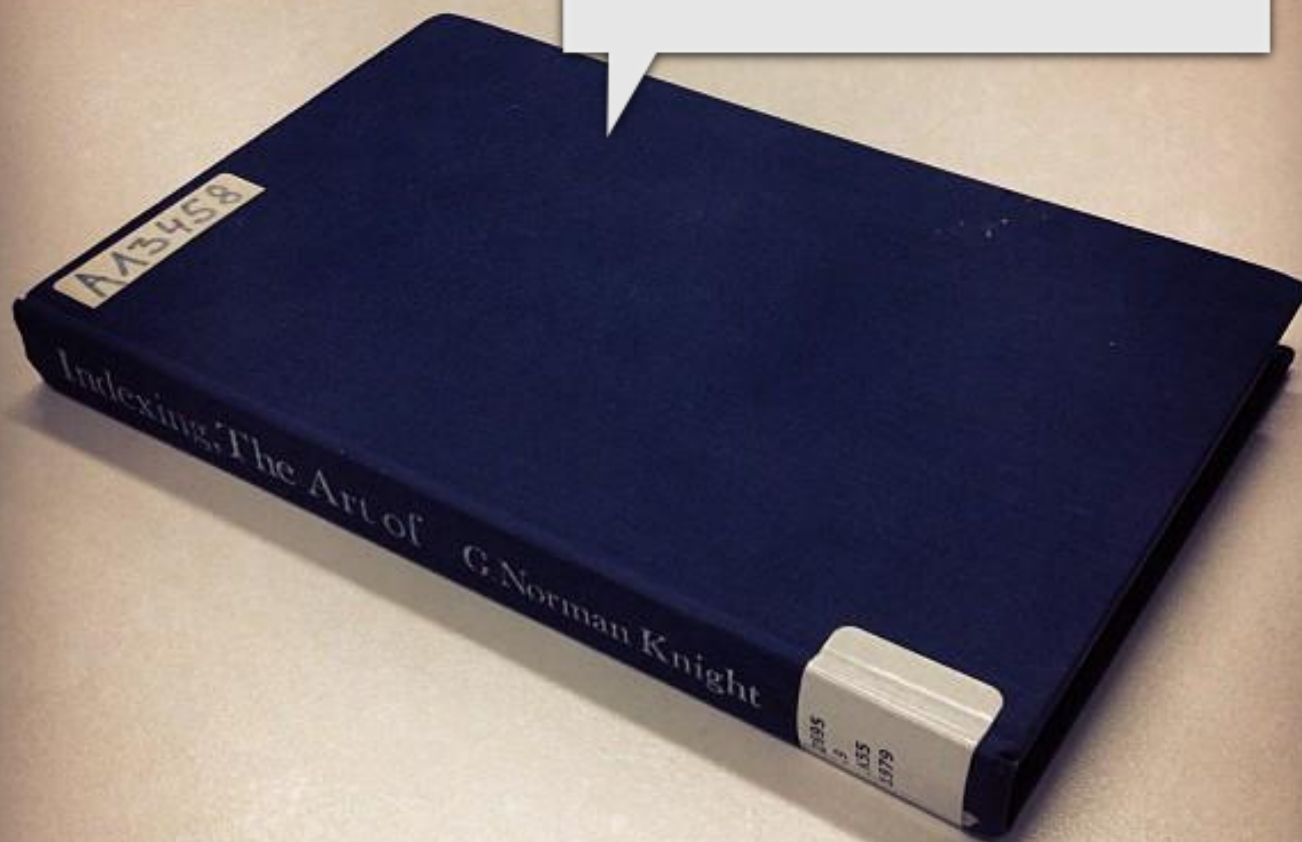
indexace?

věcný popis

“Indexing is the practise establishes correspondences between a set of key-words or index terms and individual documents or sections thereof.”

Sumarizace

Plný text



reference(s), *cont.*
"strings" of, avoidance of, 104-6,
164

two or more, to a single page,
107-8; *see also* columns, *above*
reference notes, *see* notes
Register of Companies (Board of Trade),
49

register of indexers (Society of In-
dexers), 190
of American Society, 193

rejected headings, 93

"see" cross-reference from, *see under*
"see"

"Rev.", in personal name headings, 66
reviews (of books), *see* book reviews
Richardson, Samuel, index to novels
by, 20

Richter, Anne, and American Society
of Indexers, 193

Ross, Bev Anne, and American Society
of Indexers, 193

Rothman, Dr John
and American Society
193

science fiction magazines, indexes of,
21

scientific indexing

authors and subjects separately
treated in, 29n.

letter-by-letter system in, 122

names in notes etc. in, 63

"strings" of references in author
indexes in, 105

sub-subheadings in, 56

Scobell, Henry, index to his *Act
Ordinances of . . . Parliament*

Searight, Brigadier E. E. G.
indexing of names, 77.

"see" (*see also* cross-refe-
110-2

from co-ordinates,

in cumulative in-
("discarded e

downward, fr

lar, 110-1

from obs

index

from r

from

Rejstřík

Přidat do mé schránky

Zobrazení plného záznamu:

Vyberte formát zobrazení: Standardní formát Katalogizační listek Citace

Záznam 5 z 5

Sysno

Hlavní záhlaví

Název

Naklad. údaje

Popis (rozsah)

Heslo PSH

LCC

ISBN

Jednotky

000

• B

• T

Ne

XXII, 233 s.

• věcný popis -- in

• zpracování -- ob

Z695.9 .B66 1994

ISBN 0-471-01449-4

Všechny jednotky

Zvolte formát: Standardní formát Katalogizační listek Citace S návrh

miace - Nastavení - Otázky/připomínky

Hesla PSH



Kategorie

Název

INDEXING, THE ART OF

A guide to the indexing of books and periodicals

by
G. NORMAN KNIGHT, MA
*Barrister-at-Law
Late President of the Society of Indexers*

With a Foreword by the
Rt Hon. HAROLD MACMILLAN, OM, PC, FRS

*Absente auxilio perquirimus undique frustra,
Sed nobis ingens indicis auxilium est.*

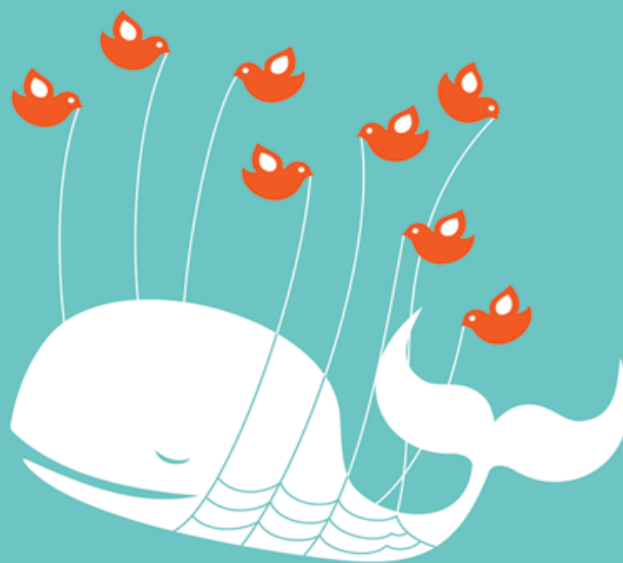
Without a key we search and search in vain,
But a good index is a monstrous gain.

LONDON
MACMILLAN & CO. LTD.

1177

218

Motivace



Zahlčení informacemi



**Škálovatelnost
indexace**



Vyhledávání a prohlížení



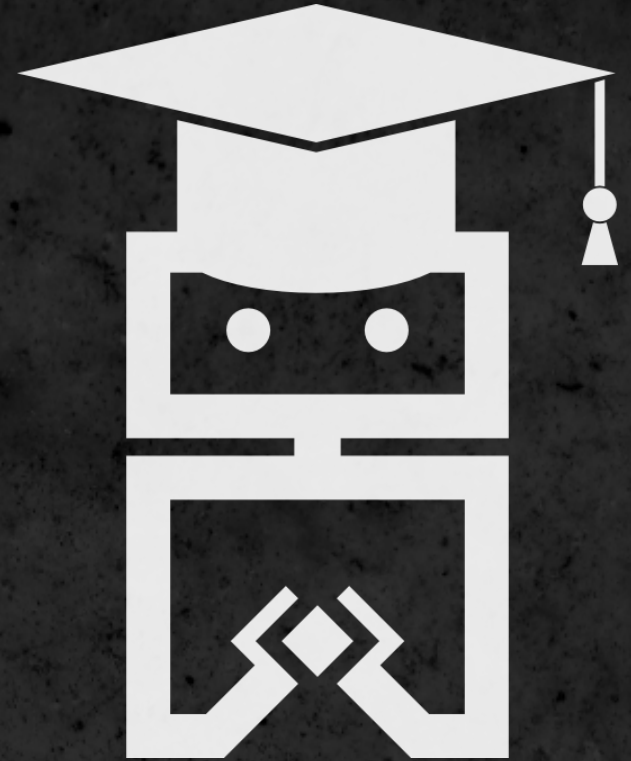
Automatická indexace

“With access to domain and general semantic knowledge, computers can index as well as humans.”

– Olena Medelyan

Strojové učení

- Učení **korespondencí** mezi rysy dokumentu a předmětovými hesly
- Kopírování a aplikace odpozorovaných pravidel



Zdroje automatické indexace

1. analýza indexovaného **dokumentu**
2. analýza použitého **řízeného slovníku**
3. analýza **způsobu použití daného slovníku** nad korpusem dokumentů

Výsledkem strojového učení nad zdroji automatické indexace je **indexační model**.

Komponenty

Indexátor



Maui Indexer

[http://code.google.
com/p/maui-indexer/](http://code.google.com/p/maui-indexer/)

Korpus plných textů

NU  **Národní úložiště**
ŠL  **šedé literatury**

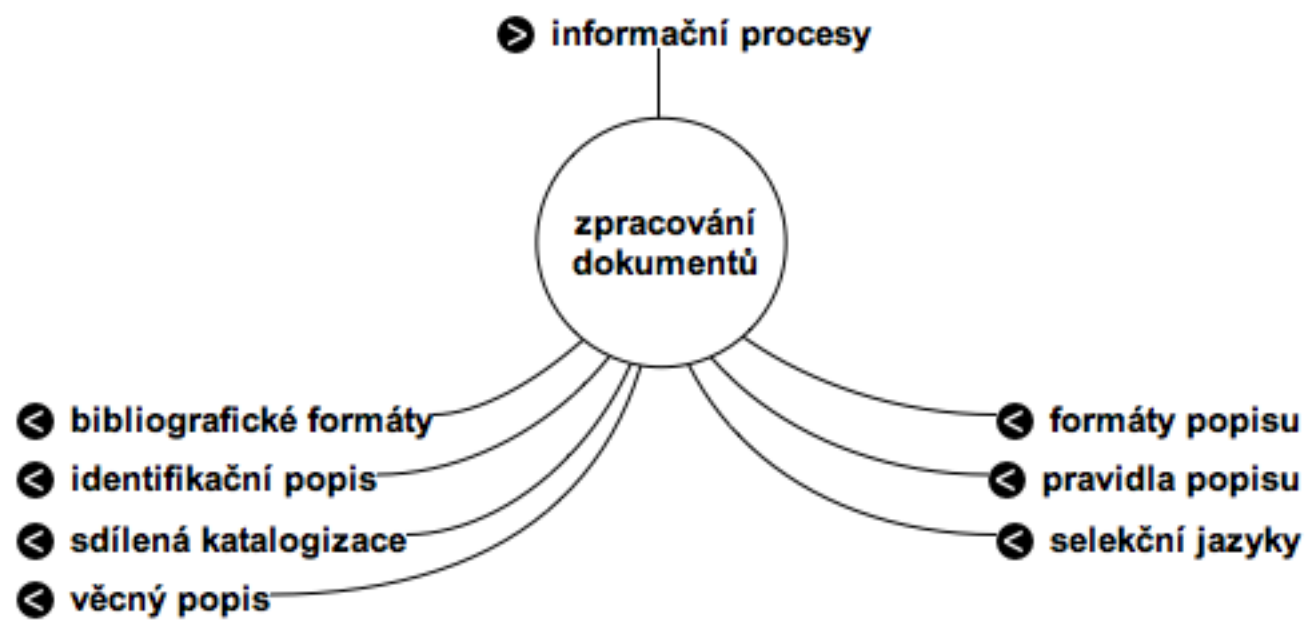
<http://nusi.techlib.cz/>

Řízený slovník

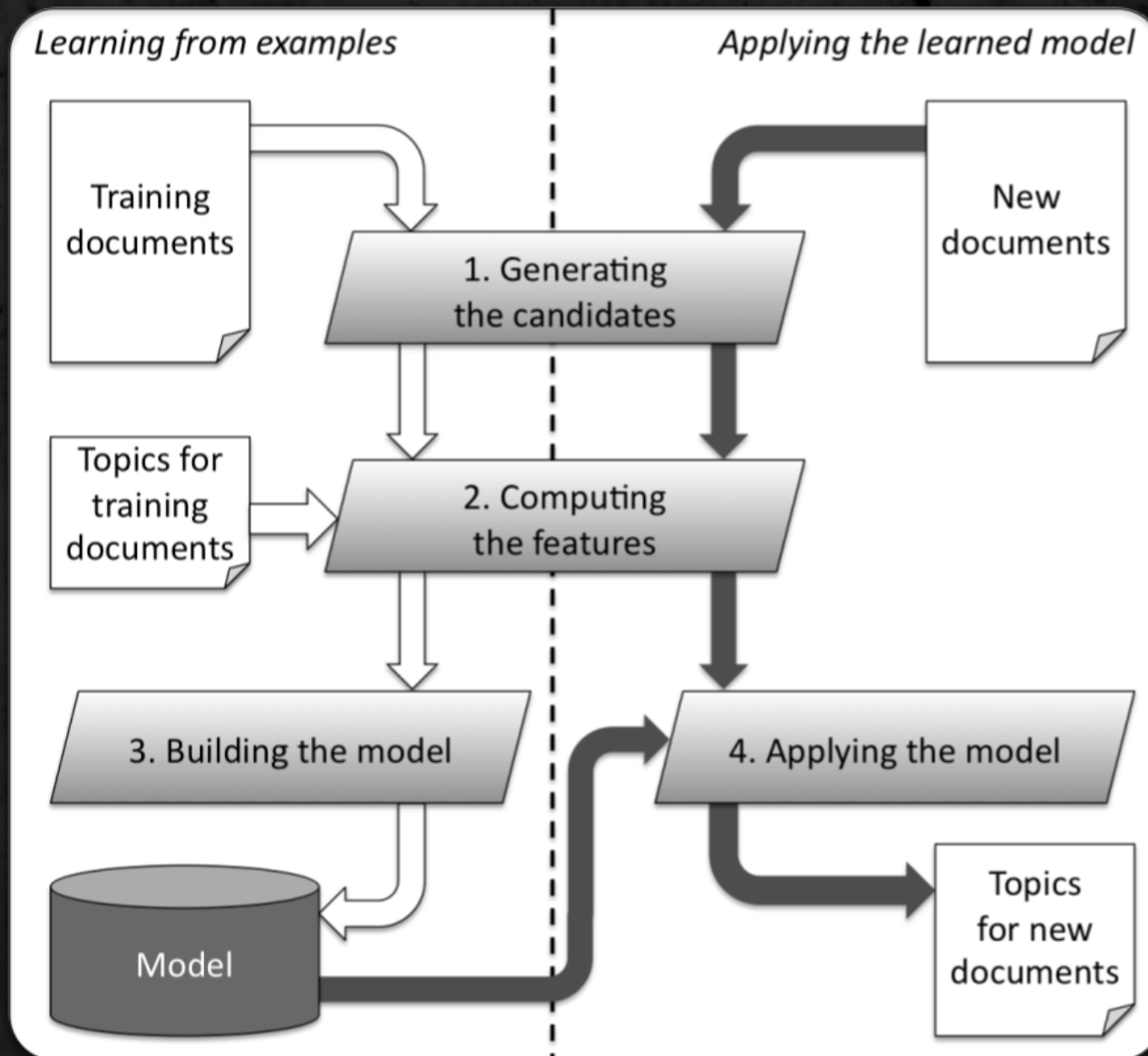


**Polytematický
strukturovaný heslář**

<http://psh.techlib.cz/skos/>



Automatické přiřazování hesel PSH

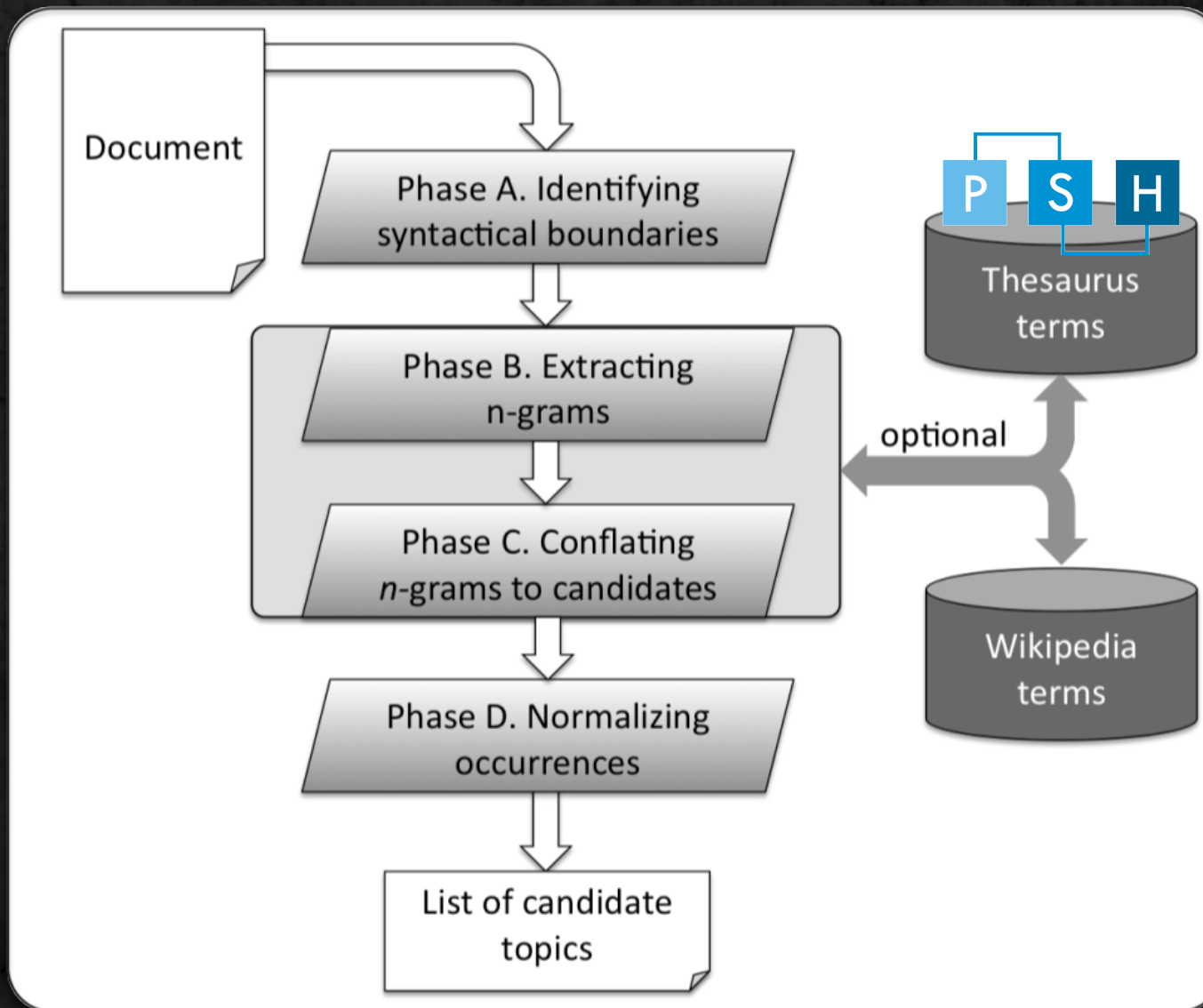


Předzpracování dokumentu

- přizpůsobeno češtině
- normalizace
- odstranění nevýznamových výrazů
 - založeny na nejfrekventovanějších výrazech v Českém národním korpusu FF UK
- redukce na slovní kořeny
 - dostupné české stemmery jsou příliš agresivní

Generování kandidátů





operations is "liable to continuous error", ^{1/} while Baxendale
"Thus far the role of the computer is chiefly that of research
it can fully assume the task of indexing is still in doubt". ^{2/}

1.2 Scope of This Study

In view of the continuing controversy over the feasibility of
indexing techniques, a state-of-the-art survey and report is pre-
time. The topic is controversial on at least five grounds: First,
indexing be done by machine at all?" Next, "Is what can be done
termed 'abstracting', 'indexing', or 'classifying'?" The third
can be done by machine good enough, acceptable, as good as,
of human operations?" The fourth and most critical question is
acceptability or comparability for any indexing process whatsoever
by man or by machine or by machine-aided manual operations

Filtrování kandidátů



Demo

<http://invenio2.ntkcz.cz/indexer/>

<http://maui-indexer.appspot.com/>

Vložte text k indexaci

Indexovat

nebo

Vložte soubor s textem k indexaci (.pdf, .txt)

Choose File No file chosen

Nahrát soubor

O projektu

Na této stránce si můžete vyzkoušet automatickou indexaci **českého** textu předmětovými hesly.

Po vložení textu (souboru s textem) dochází ke spuštění automatického indexeru ([Maui Indexer](#)), který text analyzuje, a na základě "natrénovaného" modelu se pokusí přiřadit nejvhodnější hesla vystihující jeho věcnou podstatu. Hesla jsou vybírána z [Polytematického strukturovaného hesláře](#), řízeného slovníku, který je používán v [Národní technické knihovně](#) k indexaci dokumentů.

Submit New Record

Konferenční materiály

Submit New Record

1

end of action

PSH Search

Přiřazená hesla

**automaticky
navržená hesla**

- informační věda ✕
- automatizace knihoven ✕
- software ✕
- algoritmizace ✕

Dokumenty označené heslem 'informační věda' (celkem 65 dokumentu):

[Polythematic Structured Subject Heading System](#)

[Do knihovny skrze webový prohlížeč](#)

[Using Environmental Grey Literature to Engage Public Participation in Decision Making: Government Transparency in Coastal Policy Development](#)

[Projekt NUŠL a další projekty v ČR](#)

[Integration of an Automatic Indexing](#)

[Portály v oboru Informační a knihovnictví](#)

[Current Status and Issues in Collection](#)

[From OpenSIGLE to OpenGrey : Character](#)

[Digitální knihovny - včera, dnes a zítra](#)

[Polytematický strukturovaný heslář](#)

**příklady dokumentů
s daným heslem**

[Literature Repository](#)

[Reports in Korea: Improvement through the use of the National Science and Technology](#)

inform

informační systémy pro cestující
informace
informační věda
informační prameny
informační fondy
informační procesy
informační služby
informační potřeby
informační etika
informační systémy
informační sítě
informační pracoviště
informační střediska
informatizace
informatika
informatické vzdělání
informování
informační společnost
informační management
informační zdroje
informační chování
informační systémy mezinárodní
informační systémy oborové
informační míry
informační technologie
informační bezpečnost

**automatické
dokončování**

'informační věda' (celkem 65 dokumentu):

System

Engage Public Participation in Decision Making: Government Transparency in Coastal Policy Development

em within the Document Flow of a Grey Literature Repository

and Dissemination of National Technical Reports in Korea: Improvement through the use of the National Science and Technology and Continuity

Výzvy pro indexaci

- Indexování vytvářené uživateli
- Využití **herních prvků** (<http://www.librarything.com/coverguess>)
- Efektivní využití v **uživatelských rozhraních**



<<mailto:jindrich.mynarz@techlib.cz>>

<<mailto:ctibor.skuta@techlib.cz>>

<<mailto:tomas.muller@techlib.cz>>

Ikony převzaty z <http://thenounproject.com/>.